The M4 competition: Conclusions

Spyros Makridakis[1] and Fotios Petropoulos[2]

[1] Institute for the Future (IFF), University of Nicosia, Cyprus
[2] School of Management, University of Bath, UK

There have been exactly 40 years since the publication of the predecessor of the M Competitions, the Makridakis and Hibon (MH) study (1979), with its surprising and highly contested finding that simple statistical methods were at least as accurate as complex and statistically sophisticated ones. More specifically, this study found that the Mean Absolute Percentage Error (MAPE) of Naïve 2 was, on average, 1.4% more accurate than that of the Box-Jenkins methodology to ARIMA models, while Single Exponential Smoothing (SES) was correspondingly 2.61% more accurate. Clearly, this was unwelcome evidence for the most popular approach to statistical forecasting at that time and was not received kindly by the statistical academic community (see Hyndman's paper, this issue, "A brief history of forecasting competitions" for details). Nevertheless, the MH study started a long forecasting winter that continued with the M1 (Makridakis et al., 1982) and M2 Competitions (Makridakis et al., 1993) that provided similar findings. Such findings were partially lifted with the M3 Competition (Makridakis et al., 2000), when the sMAPE of Theta, still a simple method, was 2.46% more accurate than Naïve 2, 1.31% more accurate than SES and 1% more accurate than the Box-Jenkins method.

The forecasting spring began with the M4 Competition where the top method, a complex hybrid approach combining statistical and ML elements, came first, providing an improvement in its sMAPE (symmetric MAPE) over that of the Comb benchmark by 9.4%, while the top sixteen methods achieved improvements in sMAPEs that were on average 4.49% more accurate than that of the Comb benchmark. As Gilliland (this issue) pointed out in his discussion paper, with such results, the time has come to discard the belief that complex methods are no better than simple ones, also adding that it is time for the death knell for ARIMA models, still in use today despite the empirical evidence piling for the last 40 years. Spring brings considerable opportunities that can be exploited to advance the theory of forecasting and attract more practitioners in search of ways to improve the accuracy of their predictions and more realistically assess future uncertainty. It is the purpose of this concluding paper to describe the factors that have brought the forecasting spring, briefly reiterate the major achievements of the M Competitions overall and the M4 in particular, and discuss how such factors can be exploited to avoid another winter.

Factors that brought the forecasting spring

A combination of factors that brought the forecasting spring are discussed below according to their perceived importance.

A new breed of forecasters: As the field of forecasting has been advancing, it has attracted a new breed of forecasters studying in a number of academic institutions specializing in this field, while at the same time, full-time forecasting experts are being hired by business firms interested in improving their predictions and better understanding the uncertainty involved. The major difference between the old bunch of forecasters like Spyros and the new ones like Fotios is that Spyros studied statistics, among other things, and specialized in forecasting afterwards, but has been also working on several other projects concurrently. Fotios, on the other hand, has been exclusively involved in forecasting while his PhD thesis was also in forecasting. Pablo of the second winning method has a career that is similar to that of Fotios. He completed his dissertation on forecasting with his advisors at Monash University who were statisticians that specialized, like Spyros, in forecasting later in their careers. Pablo's publications are entirely focused on forecasting as is his thesis. The same is true with Evangellos, the co-organizer of the M4 Competition and co-author of the M4 and several other papers included in this special issue. His PhD was on forecasting and he immediately started his research in this field. Practically all his numerous publications are in the forecasting field. Furthermore, Fotios, Pablo and Evangelos grew up with computers and are sophisticated programmers while Spyros started when punched cards were still used to communicate with computers.

On the practitioners' side, Slawek of the first winning method holds an MSc in Physics and has been working full time for Uber doing time series forecasting exclusively. Marciej of the third winning method also holds an MSc in computer science and is interested in neural networks and natural language processing. Slawek and Marciej, like Fotios, Pablo and Evangelos, are computer experts too, having written their own, advanced programs for their winning methods. Slawek, Marciej and the other practitioners who have contributed papers for this special issue are also working full time in forecasting and hold advanced degrees, including PhDs and Masters. Thus, the difference between the old and new group of forecasters is fundamental, which explains their significant contribution that in our view is the single most important factor for the huge accuracy/uncertainty improvements achieved in the M4 Competition. No doubt, they and other similar contributors are those that brought the forecasting spring. The forecasting field has gone from being a part-time career for a group of people interested in this field, to a full-time occupation for both academics and practitioners alike. Clearly, the last decade has witnessed the emergence of the professional, full time forecaster both in the academic and business fields.

Great advances in ML methods: Luckily, the statistical field of forecasting discovered a close cousin from the ML area, also interested in pattern recognition and, therefore, forecasting. Such a discovery offers a significant prospect to further advance the field and create an expanded, unified forecasting space. As Januschowski and colleagues (this issue) stated these two fields *"can learn and benefit from each other's strengths"* as long as they can accept *"to step outside their comfort zone"* and work together. In our view, this unification will need to happen in the near future as the complementarity of the two schools of thought is better understood and their advantages and drawbacks become clearer. There is no reason that this artificial distinction should continue to exist in the future, whereas breaking it will allow for a

wider choice to satisfy various forecasting needs, without any concern whether they are statistical or ML, as long as they can best satisfy specific forecasting applications, achieving the highest performance at the lowest cost. Slawek's wining method is the best example of the advantages to be gained by marrying the two approaches.

Exponential improvements in computer speed and memory: It would have been practically impossible to run the M4 Competition with its 100,000 series even ten years ago, or have some of the participating methods, which require enormous amounts of CPU time, to be part of the competition. No doubt, major improvements in computer speed have made the M4 Competition possible, especially for those methods utilizing complex machine learning (ML) that require demanding computations. Equally, faster computers allowed participants to experiment with various options to be able to select the best method(s) to be used in the M4 Competition (a participant told us of his huge electricity bill from running his five home computers almost constantly over 4.5 months in order to choose the best method(s)). As improvements in computer speed continue, two things will happen. First, forecasters will be able to experiment with a wider range of options to improve accuracy and uncertainty and secondly, the use of ML methods for forecasting purposes will increase as computer costs decrease and greater experience from using them will be gained.

The contributions of the M Competitions and the achievements of the M4

Hyndman's paper (this issue) provides an excellent discussion about forecasting competitions making it possible for us to be brief in this section's paragraphs. The major advantages of the M Competitions are their openness and objectivity. Their main contribution to the field and what distinguishes them from others, is their strong emphasis on **learning** and using such learning to improve the theory and practice of forecasting. The detailed results of the M Competitions have been published in respectable journals. Moreover, they have generated considerable interest in the academic community that has utilized this data to experiment with new methods. Practitioners have also been interested in these results in their search to improve their firms' forecasting performance and estimate uncertainty more realistically. Finally, the M Competitions have provided a historical record extending 40 years back that can be used in various ways to experiment and advance the field.

The four key achievements of the M4 Competition: In our view and that of the organizers, its four key achievements are as follows:

- It proved that complex methods provided accuracies that were well above that of simple ones. More specifically, sixteen methods achieved higher accuracies that the Comb benchmark, while the top six exceeded such a benchmark by more than 5%. The improvement of 19.3% of Smyl's (this issue) winning method over Naïve 2 is impressive versus the corresponding decrease of 1.4% of Box-Jenkins found in the MH study.
- It confirmed that combining more than one method improved forecasting accuracy considerably, while it also established another form of even more accurate, hybrid combining of statistical **and** ML approaches.

- It demonstrated that the two top methods provided a phenomenal precise estimation of uncertainty (as far as we know this is the first time that methods have done so) that can be utilized by other methods to estimate uncertainty realistically.
- Almost half of the submitted methods, including the top nine, have been fully replicated by the organizers of the competition and their code is available for free on GitHub, along with all the data for anyone to use in their company or for conducting academic research.

Be realistic on what forecasting can and cannot do and avoid "overselling" its achievements

The forecasting market is huge and, in our opinion, still virgin. Most business people do not believe or trust forecasting, some equating it with fortune telling and others with economic predictions that often go wrong. What the M4 Competition has shown is that forecasting accuracy can be considerably improved over the naïve approaches that are widely used in businesses. Additionally, there is a large choice of alternative methods that can be adapted to all business needs, given some specific forecasting budget and some amount of effort to be devoted to forecasting. This means that an educational effort is required to explain the advantages of systematic forecasting, but also to point out that all forecasts are inaccurate with the only certainty the extent of such inaccuracy that should be reported together with point estimates. It will therefore be necessary to assess the potential benefits from the improved accuracy and translate them into dollars and cents to persuade prospective users of the benefits of systematic forecasting for their scheduling, planning and strategic tasks. Additionally, the value of estimating uncertainty realistically must be made clear, as well as its benefits in reducing inventory cost and improving customer satisfaction among other gains.

Gilliland (this issue) in his discussion paper talks *about "the shocking disappointment of real-life business forecasting"* that falls well short of its theoretical potential, suggesting that research is needed to determine the causes that stop firms from using more accurate methods and what can be done to persuade business people of the benefits of using more appropriate forecasting techniques. He suggests using a Forecast Value Added (FVA) analysis to identify bad practices in order to avoid them and determine the potential savings from using better methods. Now that big tech firms, including Amazon, Google, Microsoft and SAP among others, are offering ML forecasting services, it will probably open the business forecasting market, thus increasing the attractiveness of forecasting as a field and creating substantial opportunities for research to further improve its value, achieving more accurate predictions. Moreover, business firms like Uber use forecasting heavily for their day-to-day operations, highlighting its benefits and becoming an example for other firms to follow.

Closing remarks

This Special M4 Competition Issue has brought together high caliber academics and top-level practitioners to comment, discuss and criticize the M4 Competition, as well as the challenges facing the field and the best way forward. Their suggestions have been invaluable and have established a direct line of communication between the academic and business communities

that we hope will grow and strengthen in the future. Similarly, we expect that the integration of the statistical and ML groups will be successful, providing a common effort to further advance the field. As a part of the wider data science field, forecasting is bound to play a critical future role in identifying patterns in data and consequently forecasting as accurately as possible, while also realistically estimating uncertainty. We should add that this special issue contains a detailed description of the winning methods by their authors that can serve as the starting point for using them, conducting additional academic research, and improve the practice of forecasting in business firms while making sure that practitioners fully understand its benefits but also its limitations and the fact that all future predictions are uncertain.

Once this issue is finalized, planning for the M5 Competition will accelerate. Its major difference from the M4 will be the inclusion of explanatory and exogenous variables to determine if doing so will improve the accuracy of pure time series predictions. As in the M4 Competition, several difficult choices will have to be made. These will be done by taking into account the feedback and numerous suggestions received from the forecasting community. Hopefully progress will continue and future competitions will further improve forecasting accuracy/uncertainty and contribute to persuading more business users of the advantages and practical benefits of utilizing systemic predictions as an integral part for scheduling and planning as well as formulating strategies. Continuous progress and realism of what forecasting can and cannot do will in our view will avoid another forecasting winter.

References

Gilliland, M. (2019). The Value Added by Machine Learning Approaches in Forecasting. International Journal of Forecasting, 35(4), XXX-XXX.

Hyndman, R. (2019). A brief history of forecasting competitions. International Journal of Forecasting, 35(4), XXX-XXX.

Januschowski, T., Gasthaus, J., Flunkert, V., Wang, B., Bohlke-Schneider, M., Salinas, D., & Callot, L. (2019). Criteria for Classifying Forecasting Methods. International Journal of Forecasting, 35(4), XXX-XXX.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. Journal of Forecasting, 1(2), 111–153.

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. International Journal of Forecasting, 16(4), 451–476.

Smyl, S. (2019). Exponential Smoothing and Recurrent Neural Network Hybrid Model. International Journal of Forecasting, 35(4), XXX-XXX.